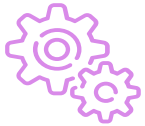
Price Oracle

end-to-end solution for price monitoring and prediction

Gianfranco Demarco - Francesco Ranieri

Supervision: Dott. Antonio Pellicani

Project objectives



Automated pipelines

Data ingestion is automated with real-time pipelines that fetch the most updated data

Price Predictions

Deep learning models are employed to automatically produce predictions for the price of almost 20 cryptocurrencies

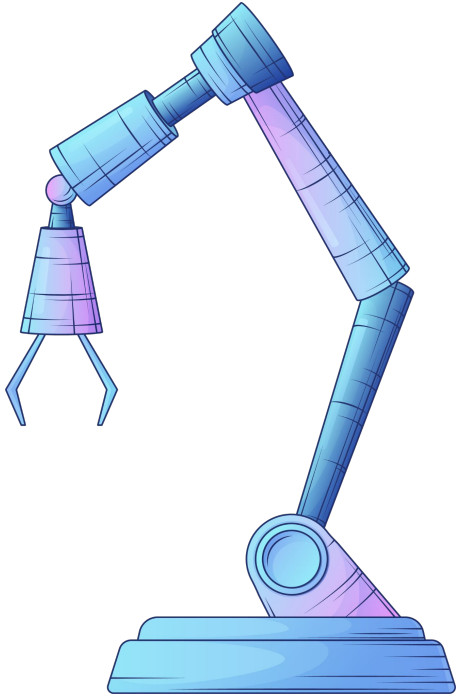
Data Visualization

Dashboards with modern layouts are provided to the user to constantly monitor all of the data in the system

Project objectives

Full automation and reproducibility

Using cutting-edge technologies such as Docker, Kubernetes and Helm, the project environment is fully automated and reproducible.



01

Data

Data - Data format

A common format to represent assets data is the **OHLCV format.**

A OHLCV data point is composed of:

- **Open:** the initial price of the asset for the reference time frame
- **High:** the highest price reached by the asset's price for the reference time frame
- **Low:** the lowest price reached by the asset's price for the reference time frame
- **Close:** the final price of the asset for the reference time frame
- **Volume:** the total quantity of the assets exchanged in the reference time frame

The time frame considered in this project is the **daily timeframe: a data point for each day is recorded.**

Data - Candlesticks

The OHLCV format is commonly used to plot **candlestick graphs.**
A sequence of OHLCV data points for an asset form a **time series.**

Data - Sources

Two sources are used for the data:

- the first is CriptoDataDownload, used to fetch the historical data of the prices
  - CriptoDataDownload offers data from a variety of exchange.
    Binance was selected as the source exchange for the data of this project.
    From the 286 cryptocurrency available, the 17 having data *at least* from 01/01/2019 were selected

- the second is Kraken, used to fetch daily updates
  - the updates are fetched for the same subset of 17 cryptocurrencies mentioned above

Of the OHLCV data, the **close price** is the **target** of the analysis.

Data - Cryptocurrencies

The 17 chosen cryptocurrencies are:

Cardano	**Binance Coin**	**Bitcoin**	VeChain	Ontology
EOS	Ethereum Classic	**Ethereum**	**Stellar**	Qtum
Icon	IOTA	Litecoin	Tron	NULS
XRP	NEO			





Data - Calculated Data

The *close price time series* are enriched by calculating the **Simple Moving Average (SMA)** at with different window sizes.
The chosen window sizes (days) are:

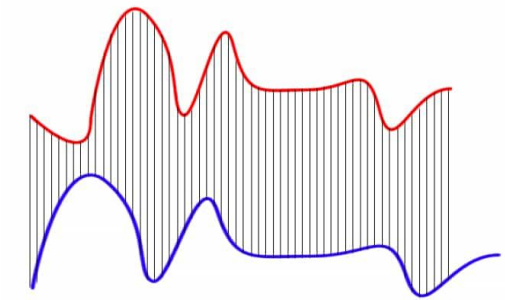
- 5
- 10
- 20
- 50
- 100
- 200

Data - Clustering

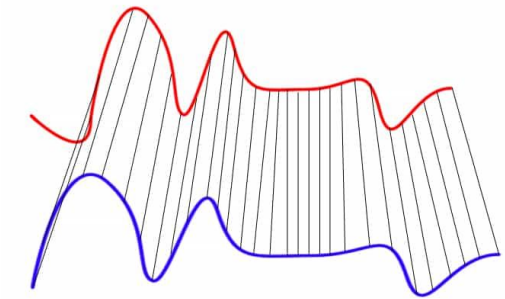
Two types of models were used: **single-target** and **multi-target.**
To identify the cryptocurrencies to group together for the multi-target models, the price time-series have been clustered.

For the purpose of clustering the cryptocurrencies, the **Dynamic Time Warping (DTW)** algorithm
was used to extract distances between the time series.
In time series analysis, DTW is an algorithm for measuring similarity between two temporal sequences.

Differently from the Euclidean distance, DTW takes into account that the two time series could be not aligned and have different speeds.



Euclidean Matching



Dynamic Time Warping Matching

Data - Clustering - Methodology

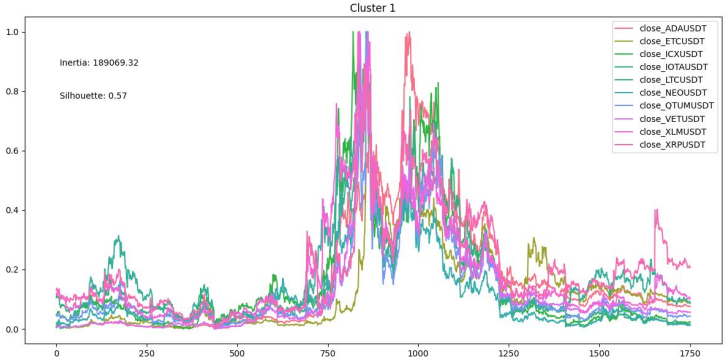
Different parameters have been used for the clustering of the time series:

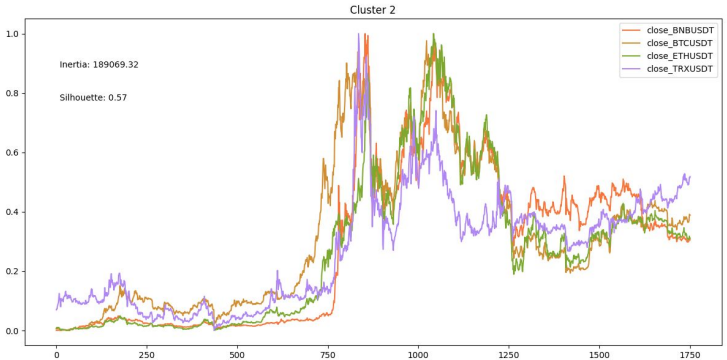
- **Clustering algorithm**: K-means or K-medoids
- **Number of clusters:** from 3 to 8
- **Data scaling**: None, Min-Max, Logarithmic
- **Percentage change**: Yes, No

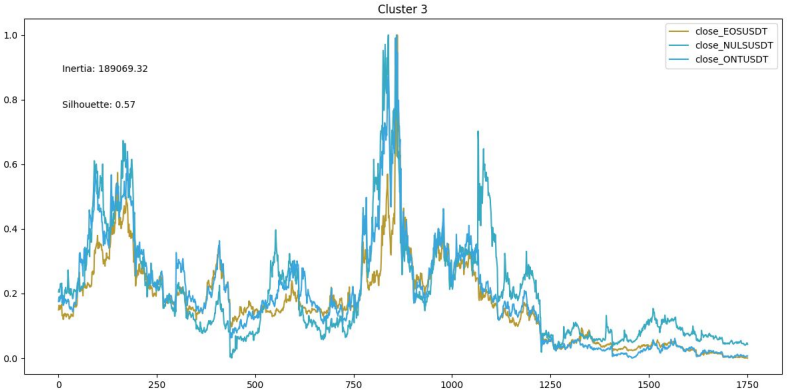
The clustering results have been evaluated using the **Silhouette Score.**

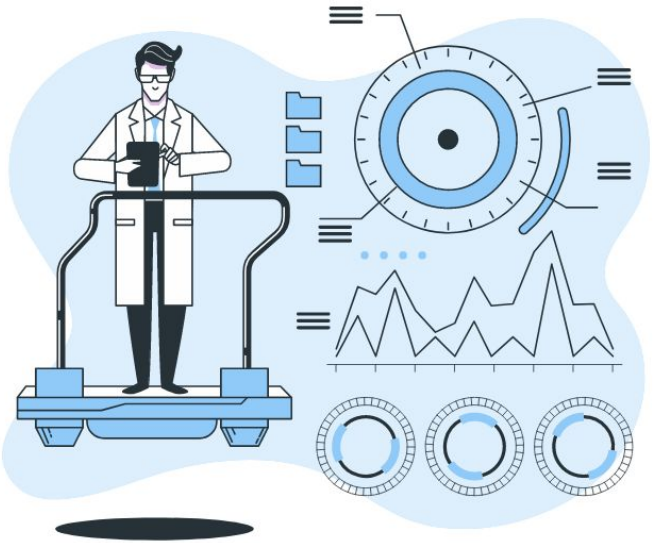
The Silhouette Score quantifies how well a data point fits into its assigned cluster and how distinct it is from other clusters.

Data - Clustering - Results









02

Modeling

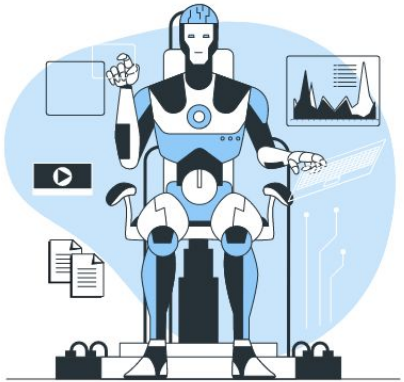
Modeling

Modeling refers to the process of *finding the best model* that reliably predicts the closing price for each cryptocurrency or cluster of cryptocurrencies.

In order to achieve this goal, different models and configurations were tested.

For better organization, management and reproducibility of experiments:

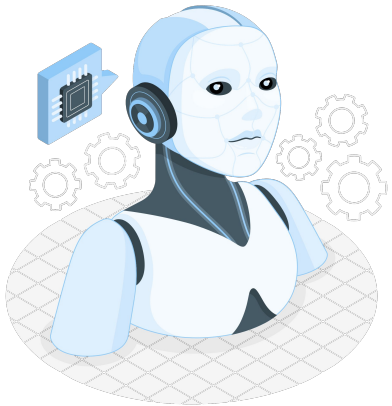
- **MlFlow** is used for experiment tracking
- **Optuna** is used for hyperparameters optimization



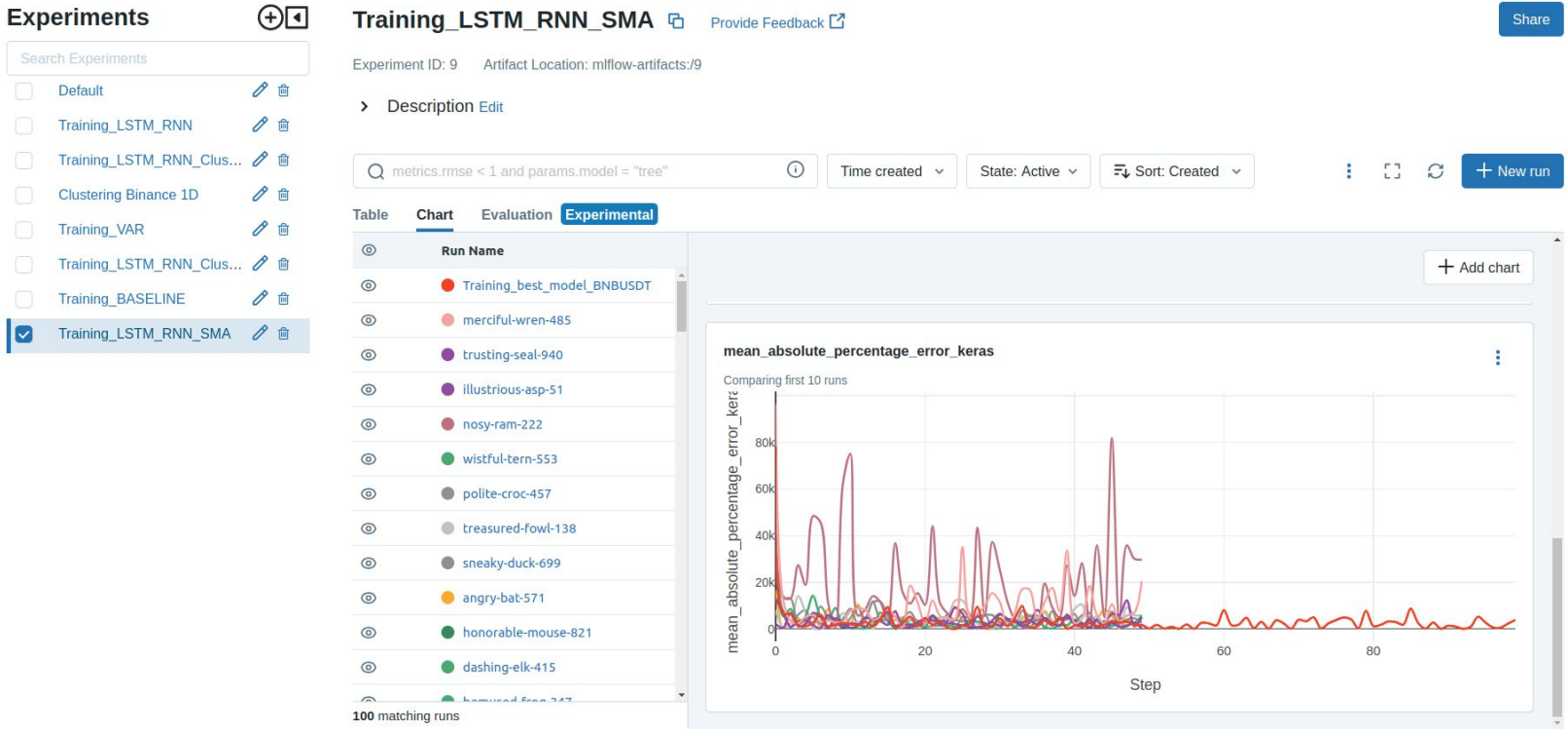
Modeling - MlFlow

MLFLow is a software that allows to track Machine Learning experiments and models.

It stores the metrics of the experiments, allowing the developer to compare different models and parameters. Also, allows to store the models and retrieve them when needed.



Modeling - MlFlow



Modeling - Optuna

Optuna is an open source hyperparameter optimization framework to automate hyperparameter search.

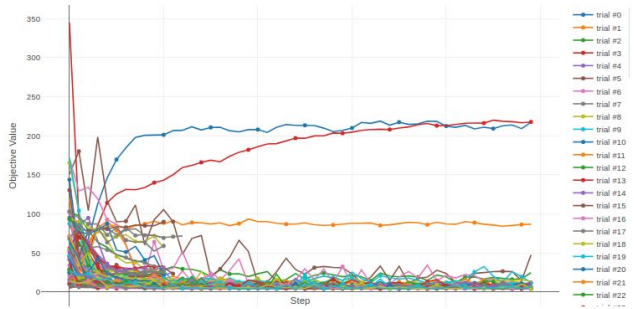
The main features of Optuna used in this project are:

- **Systematic exploration of hyper-parameters space**
- **Early Pruning**

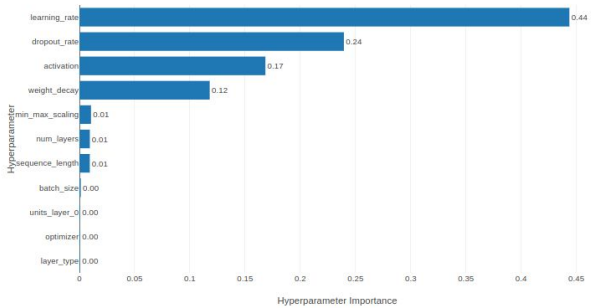
By using Optuna, 100 configurations of hyper-parameters were tested for each model.

Modeling - Optuna

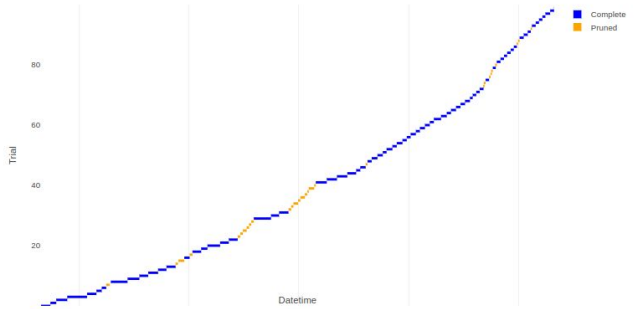
Intermediate values



Hyperparameter Importance



Timeline



Best Trial (number=95)

4.156256675720215

Params = [num_layers: 1, units_layer_0: 16, sequence_length: 4, learning_rate: 0.0008225689681198286, dropout_rate: 0.13863112552011242, min_max_scaling: 0, layer_type: RNN, optimizer: RMSprop, activation: tanh, weight_decay: 0.36963097505673276, batch_size: 32]

DETAILS

Modeling - Baseline

When predicting time series and in particular financial data, it is easy to achieve result that seem *too good to be true*.

This is because prices are **strongly related** to their previous values and next values don't deviate too much.

For this reason, a very simple *algorithmic model* was evaluated, to be used as **baseline**, against which machine learning models performance are measured.

The algorithm simply predicts, for the next day, the closing price of the previous day.

Modeling - Baseline - Results

		RMSE				MAPE			
		7 days	30 days	90 days	All	7 days	30 days	90 days	All
CLUSTER 1	ADA	2.88E-03	3.99E-03	4.96E-03	4.85E-02	0.93%	1.22%	1.38%	3.68%
	ETC	1.85E-01	2.55E-01	3.22E-01	2.30E+00	1.08%	1.31%	1.41%	3.61%
	ICX	4.30E-03	4.62E-03	5.33E-03	6.13E-02	2.05%	2.13%	2.16%	4.37%
	IOTA	3.10E-03	2.99E-03	3.81E-03	5.02E-02	1.70%	1.56%	1.74%	3.71%
	LTC	1.03E+00	1.23E+00	2.01E+00	6.77E+00	1.27%	1.47%	1.90%	3.49%
	NEO	1.20E-01	1.52E-01	1.74E-01	2.09E+00	1.43%	1.68%	1.65%	3.84%
	QTUM	2.47E-02	4.94E-02	6.07E-02	5.28E-01	1.01%	1.54%	1.90%	4.09%
	VET	1.85E-04	2.95E-04	3.94E-04	4.58E-03	0.88%	1.35%	1.62%	4.18%
	XLM	2.02E-03	1.92E-03	3.81E-03	1.38E-02	1.45%	1.40%	2.18%	3.49%
	XRP	5.40E-03	8.96E-03	1.46E-02	3.88E-02	0.89%	1.43%	1.69%	3.36%
CLUSTER 2	BNB	2.65E+00	2.48E+00	9.85E+01	1.33E+01	1.04%	0.96%	0.99%	3.17%
	BTC	5.27E+02	4.20E+02	4.56E+02	1.05E+03	1.05%	1.04%	1.00%	2.38%
	ETH	2.34E+01	2.82E+01	2.78E+01	8.26E+01	1.12%	1.23%	1.10%	3.16%
	TRX	1.20E-03	1.31E-03	1.29E-03	3.12E-03	1.17%	1.15%	1.17%	3.16%
CLUSTER 3	EOS	8.43E-03	9.52E-03	1.44E-02	2.86E-01	1.31%	1.28%	1.55%	3.63%
	NUL	3.01E-03	3.36E-03	3.82E-03	3.55E-02	1.44%	1.46%	1.54%	4.30%
	ONT	1.54E-03	3.76E-03	4.60E-03	5.68E-02	0.83%	1.70%	1.90%	4.05%

Modeling - VAR

Vector Autoregression (VAR) models are a class of time series models commonly used in econometrics and statistics to analyze and forecast *multivariate* time series data.

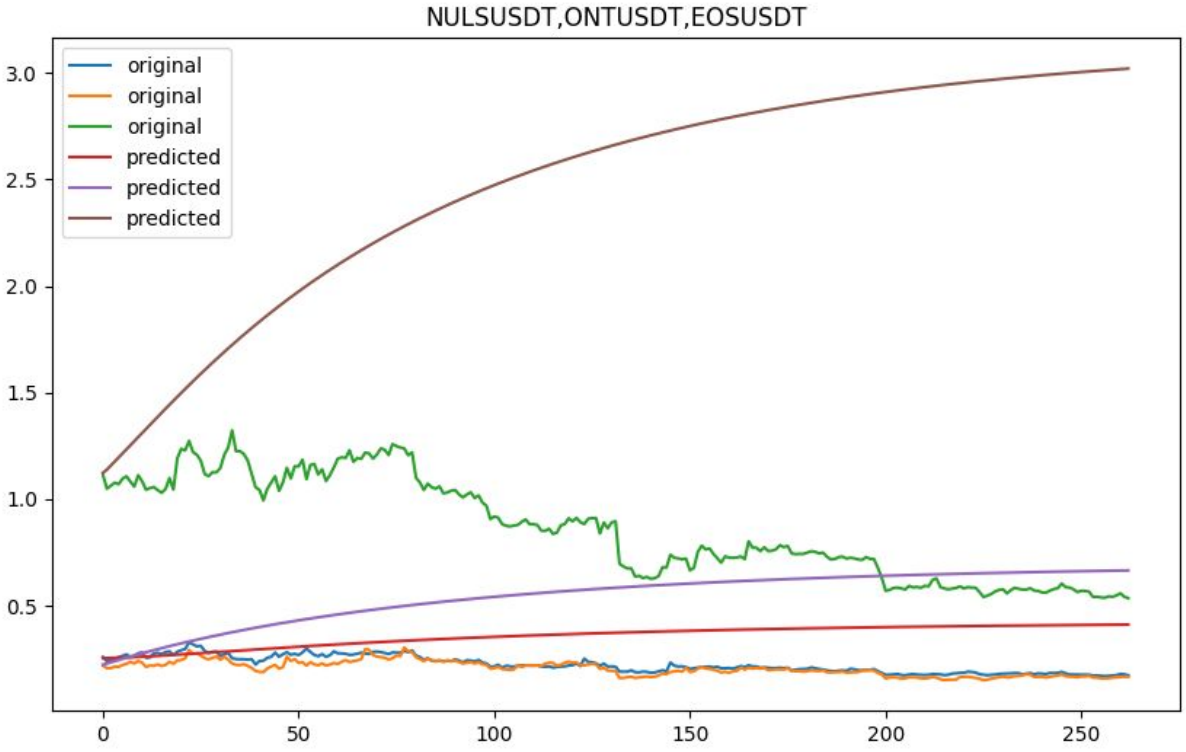
VAR models are a natural extension of univariate autoregressive models (AR), allowing for the simultaneous modeling of multiple related time series variables.

In VAR models, each variable is regressed on its own past values and the past values of all the other variables in the system. This captures the interdependencies between the variables over time.

Modeling - VAR - Results

		RMSE				MAPE			
		7 days	30 days	90 days	All	7 days	30 days	90 days	All
CLUSTER 1	ADA	1.37E+01	1.26E+01	1.04E+01	6.87E+00	130.75%	121.11%	100.11%	60.54%
	ETC								
	ICX								
	IOTA								
	LTC								
	NEO								
	QTUM								
	VET								
	XLM								
	XRP								
CLUSTER 2	BNB	4.55E+03	4.64E+03	5.57E+03	5.87E+03	42.73%	44.72%	48.67%	41.12%
	BTC								
	ETH								
	TRX								
CLUSTER 3	EOS	1.46E+00	1.44E+00	1.38E+00	1.05E+00	299.51%	285.76%	259.73%	157.63%
	NUL								
	ONT								

Modeling - VAR - Results



Modeling - Neural Networks

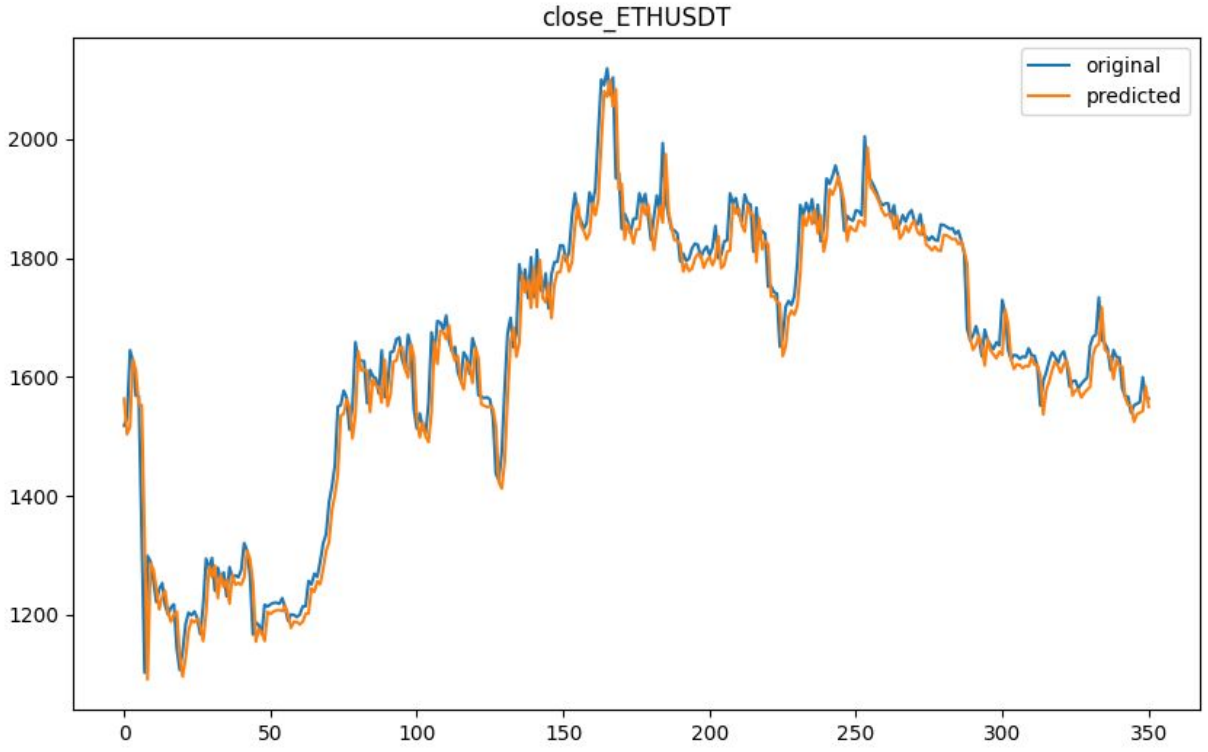
When dealing with time series, **LSTM** and **RNN** are the most common Neural Network architectures employed.

In this project, the type of network (LSTM or RNN) has been considered just as another **hyperparameter** to **optimize**, along with the number of layers, the number of neurons per layer and other hyperparameter.

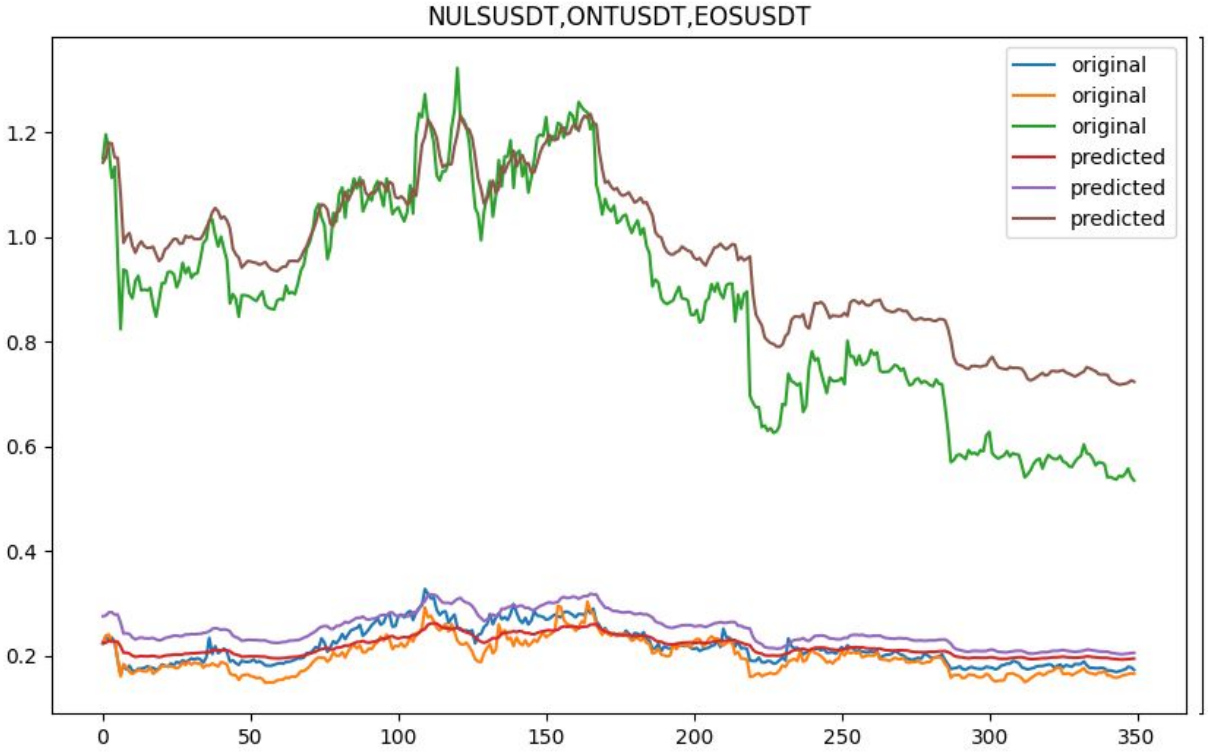
Modeling - Neural Networks - Results

		RMSE								MAPE							
		7 days	30 days	90 days	All	7 days	30 days	90 days	All	7 days	30 days	90 days	All	7 days	30 days	90 days	All
CLUSTER 1	ADA	0.003661	0.004473	0.005734	0.012100	0.661383	0.728518	0.937646	1.231554	1.23%	1.42%	1.60%	2.50%	7.95%	8.15%	8.59%	7.35%
	ETC	0.226092	0.282202	0.343965	0.723315					1.28%	1.45%	1.55%	2.52%				
	ICX	0.026535	0.027748	0.027203	0.025952					15.17%	15.84%	14.89%	12.30%				
	IOTA	0.091539	0.086602	0.076964	0.052358					63.58%	58.17%	48.48%	26.32%				
	LTC	1.819315	1.719960	2.502742	3.337073					2.65%	2.27%	2.58%	2.90%				
	NEO	0.697475	0.640178	0.561756	0.616971					10.08%	8.76%	7.07%	6.10%				
	QTUM	0.031101	0.056185	0.065994	0.106728					1.21%	1.97%	2.13%	2.79%				
	VET	0.000689	0.000756	0.000793	0.001020					4.05%	4.17%	4.25%	4.14%				
	XLM	0.005243	0.005480	0.006230	0.006760					4.69%	4.72%	4.37%	5.68%				
	XRP	0.006735	0.009968	0.014061	0.024758					1.04%	1.37%	1.81%	2.44%				
CLUSTER 2	BNB	2.826398	2.499788	3.103987	7.826777	368.146489	293.871033	298.876083	408.005913	1.07%	0.94%	1.02%	1.78%	3.01%	2.88%	2.77%	3.14%
	BTC	661.779071	725.782320	819.176291	817.259888					2.33%	2.45%	2.63%	2.74%				
	ETH	40.022632	42.053206	41.839232	56.351512					2.09%	2.18%	2.14%	2.74%				
	TRX	0.002807	0.002507	0.002348	0.002286					2.88%	2.51%	2.52%	2.73%				
CLUSTER 3	EOS	0.107796	0.107936	0.108200	0.103253	0.129314	0.124396	0.106504	0.064148	19.78%	19.08%	17.57%	12.09%	22.62%	21.69%	17.76%	9.15%
	NUL	0.005918	0.006769	0.007084	0.010304					2.98%	3.35%	3.33%	3.58%				
	ONT	0.007446	0.007500	0.007694	0.011701					4.49%	4.08%	3.98%	4.48%				

Modeling - Neural Networks - Results



Modeling - Neural Networks - Results



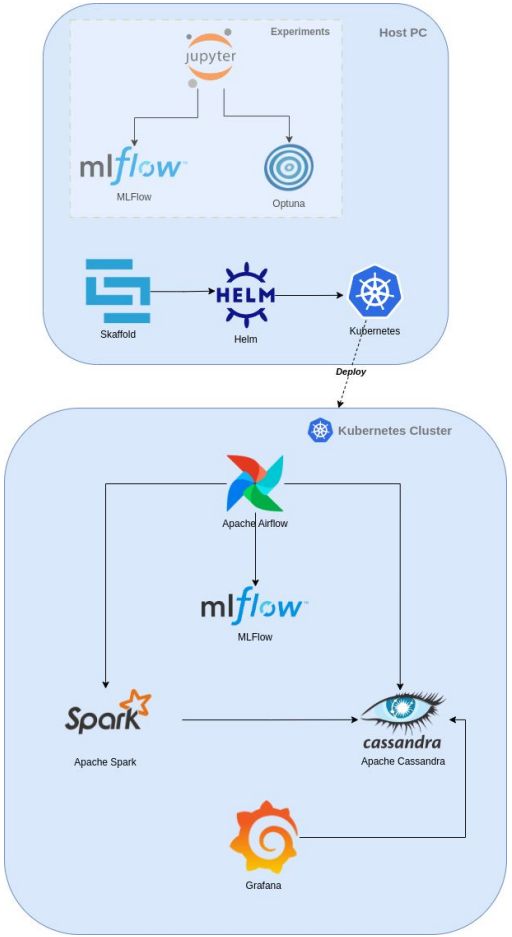
03

Architecture

Architecture

Price Oracle is developed in a **fully containerized environment**, composed of various components.

The application flow is entirely run on a **Kubernetes** cluster, which is deployed locally using Kind.
In these section the components will be briefly described.

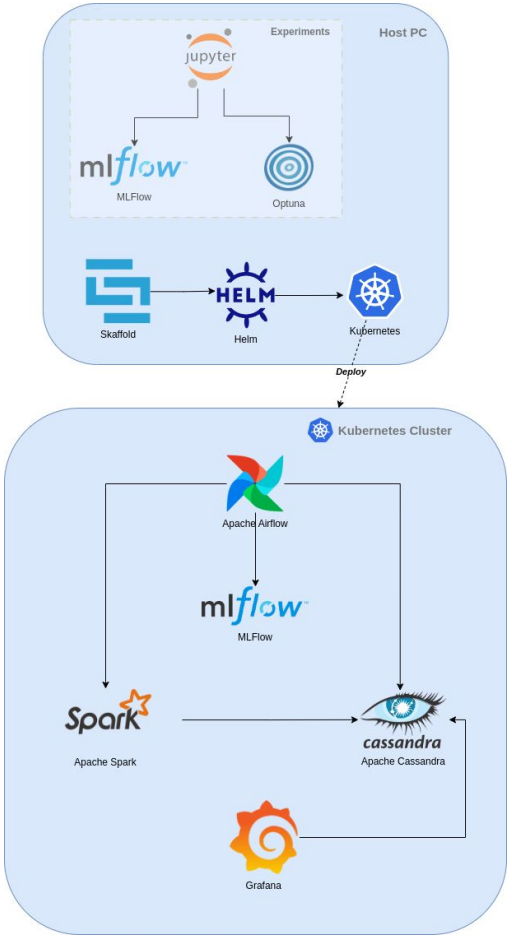


Architecture

Docker is a technology that allows to **package** an application along with all of its dependencies into a **standardized** unit for software development.

Kubernetes is an open-source **container-orchestration** system for automating computer application *deployment, scaling, and management*.

Helm is a *package manager* for Kubernetes. Helm charts are used to define, install, and upgrade Kubernetes applications.

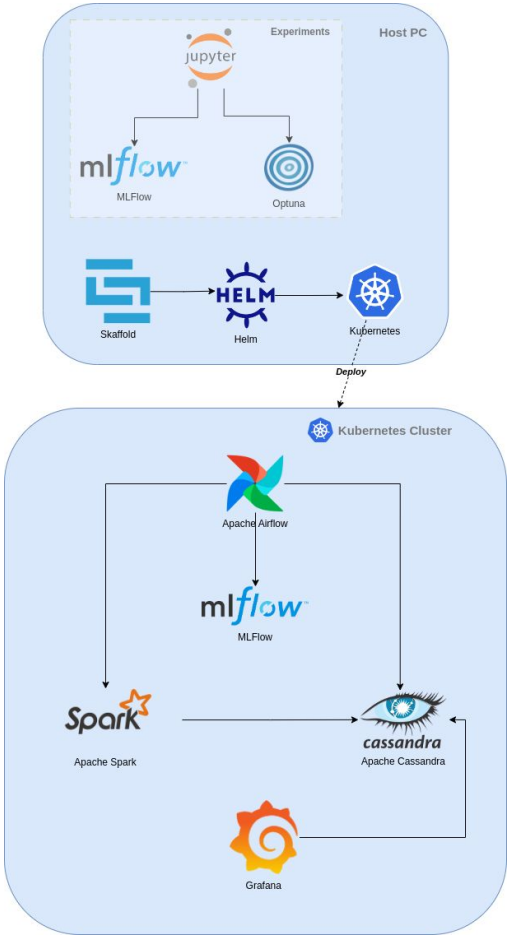


Architecture

Skaffold is a command line tool that facilitates **continuous development** for Kubernetes applications.
It handles the workflow for building, pushing and deploying applications

Kind is a tool for **running local Kubernetes** clusters using Docker container "nodes".

Apache Airflow is an open-source **workflow management** platform.
It allows to define, schedule and monitor workflows, which are defined as **DAGs** (Directed Acyclic Graphs).



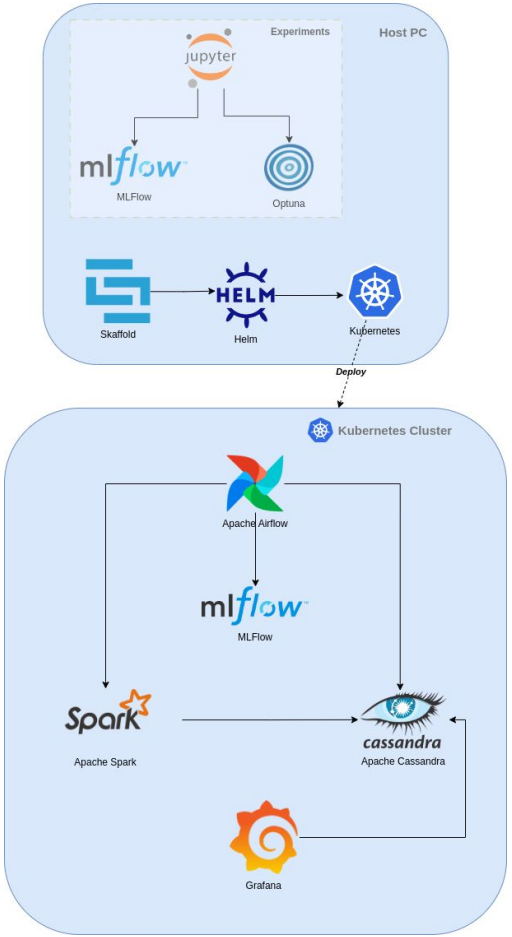
Architecture

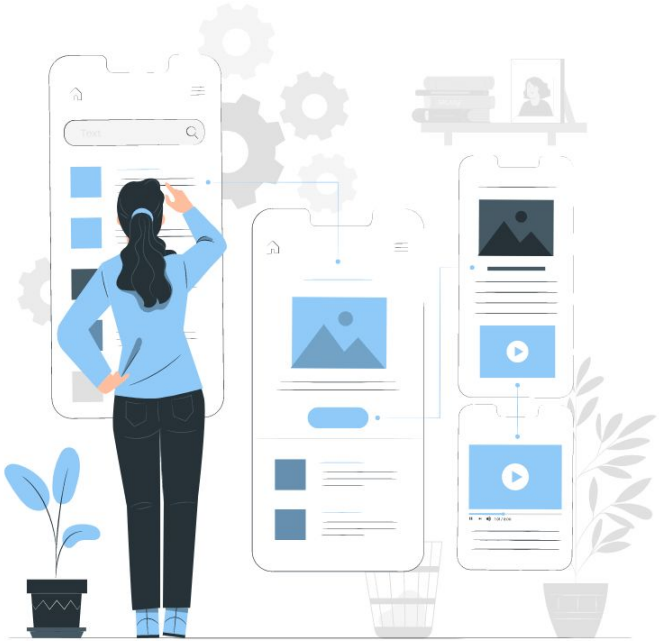
Apache Cassandra is a free and open-source, distributed, wide column store, NoSQL database management system.

Apache Spark is an open-source unified analytics engine for large-scale data processing.

Mlflow is an open-source platform to manage the ML lifecycle, including experimentation, reproducibility and deployment.

Grafana is an open-source analytics and monitoring solution.





04

Application Flow

Application Flow

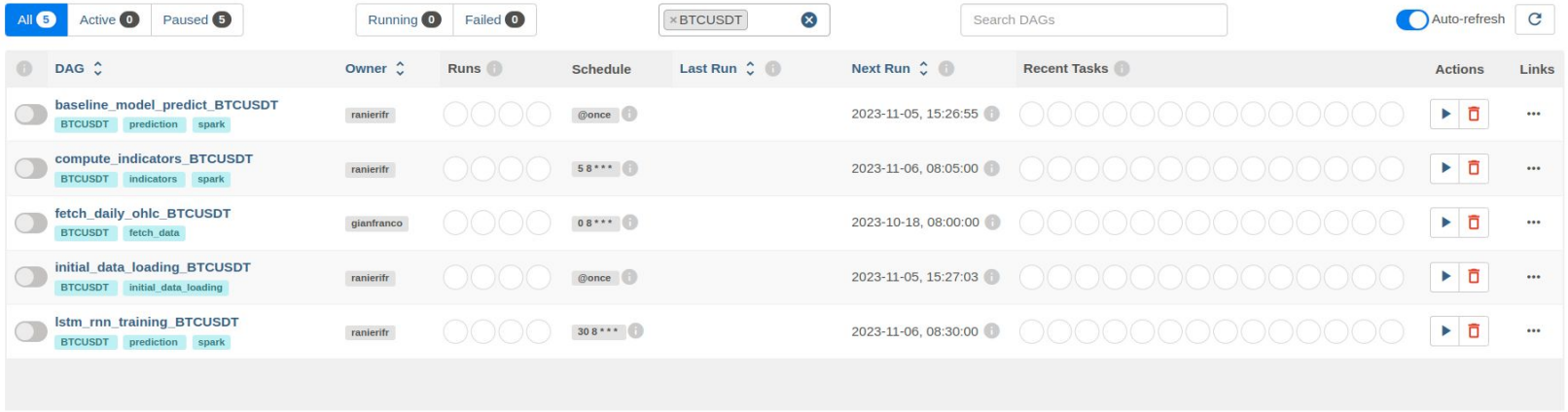
The hearth of the **Price Oracle** project are the **Airflow DAGs**, which define the entire application flow.

When the cluster is first spin up, a CSV file containing the historical data for each one of the 17 cryptocurrencies is packed into the Airflow image

The CSVs contain OHLCV data for each cryptocurrency, with the prices referring to the pair with the USDT stable coin.

For each of the cryptocurrencies, a set of 5 dags is dynamically defined, for a total of 85 DAGs.

Application Flow - Airflow - Dags

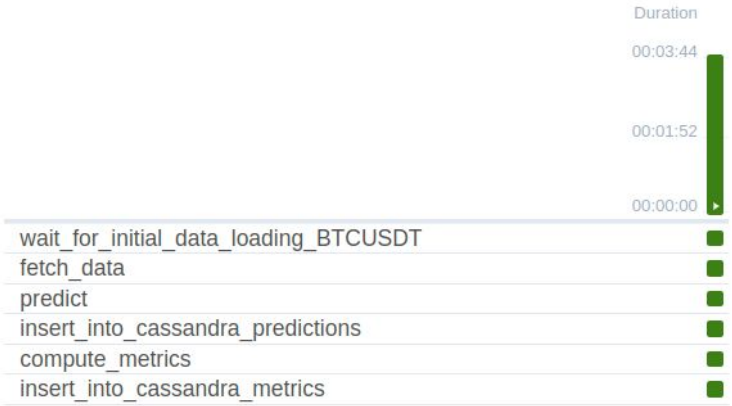


Application Flow - Airflow - Dags

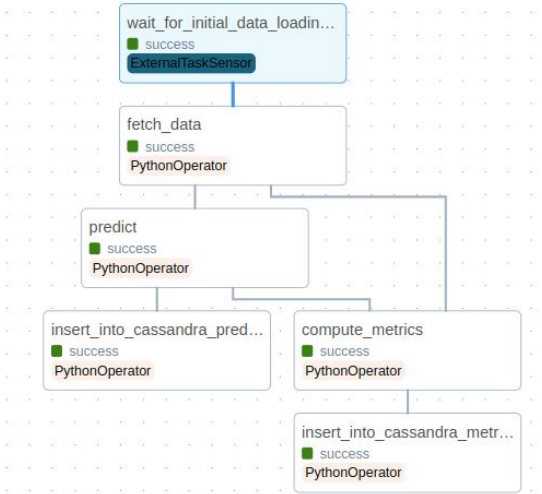
Name	Description	Scheduling
initial_data_loading	Ingest the initial data from the CSV file into the database	Once
fetch_daily_ohlc	Ingest real-time data from the Kraken API into the database	Every day 08:00 AM
compute_indicators	Periodically compute SMA indicators on the data stored in the database	Every day at 08:05 AM
baseline_model_predict	Perform predictions with a baseline model, storing the results and performance metrics in the database	Every day at 08:30 AM
lstm_rnn_training	Train a custom model, perform predictions and store the results and performance metrics in the database	Every day at 08:30 AM

Application Flow - Airflow - Dags - Example

Grid View



Graph View



05

Visualization

Visualization

Another core component of the Price Oracle project is the **Price Oracle Dashboard**, which is automatically created in **Grafana**.

The dashboard is **dynamically created** and leverages the *variables feature* of Grafana.

The list of the cryptocurrencies is stored into a Grafana variable, and this allow the user to select the cryptocurrency to visualize in the dashboard.

Visualization - Price Oracle Dashboard

BTCUSDT MAPE

MODEL NAME ↓	7 days	30 days	60 days	All
NEURAL NETWORK	0.02	0.02	0.02	0.02
BASELINE	0.02	0.04	0.04	0.03
BASELINE	0.02	0.04	0.04	0.03

BTCUSDT RMSE

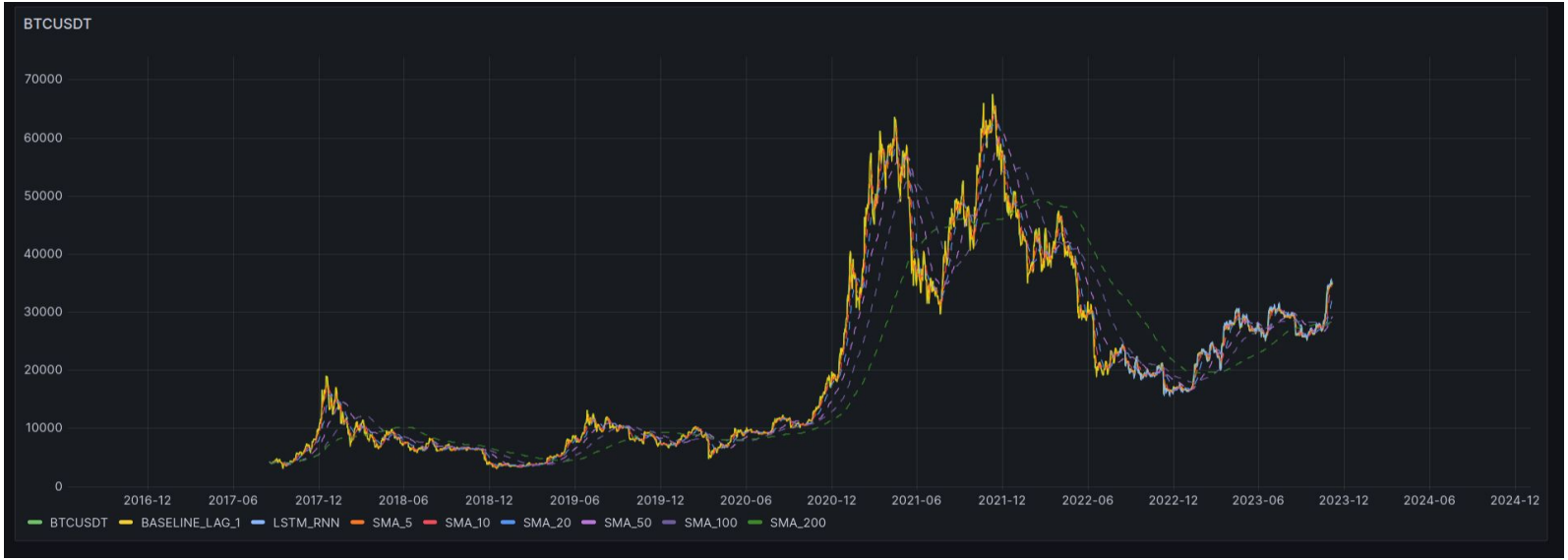
MODEL NAME ↓	7 days	30 days	60 days	All
NEURAL NETWORK	504.80	627.22	587.09	613.73
BASELINE	111.42	237.59	247.07	949.22
BASELINE	111.42	237.59	247.07	949.22

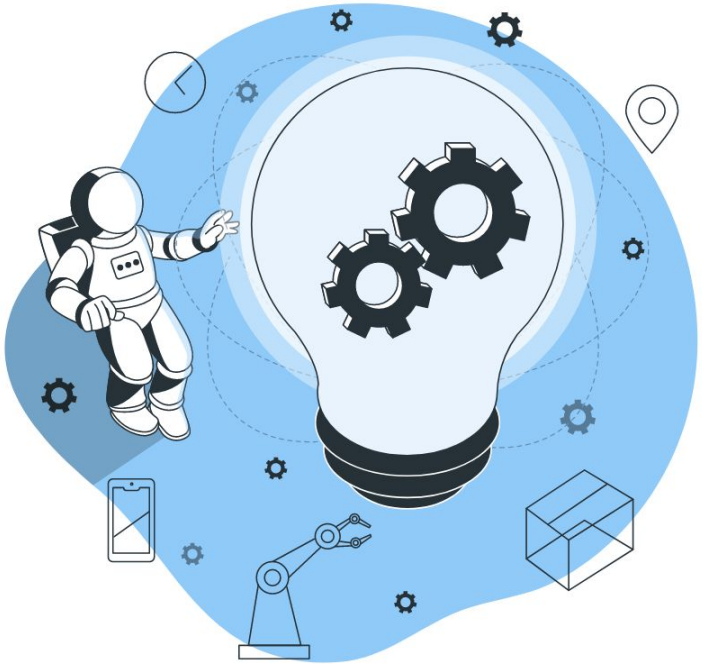
Visualization - Price Oracle Dashboard

BTCUSDT MAPE

MODEL NAME ↓	7 days	30 days	60 days	All
NEURAL NETWORK	0.02	0.02	0.02	0.02
BASELINE	0.02	0.04	0.04	0.03

Visualization - Price Oracle Dashboard





06

Future Works

Future Works

1) **Concept drift detection**

Today, the models are automatically re-trained every day.
A more thoughtful approach would be to detect a possible ***concept drift*** in the data, and retrain only in that case. An easy way to do this would be to compare the model 30 days MAPE with the model 7 days MAPE.

2) **Models training**

Only the pipelines for the LSTM and RNN models have been implemented, and only for the version considering the price time-series only.

Other pipelines can be implemented to automatically train:

- Models considering the price time-series and the SMA indicators
- Models considering the cryptocurrencies in clusters
- VAR models

Thanks!

gianfranco.demarco26@studenti.uniba.it
f.ranieri27@studenti.uniba.it